GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning

Year: 2020 | Citations: 100 | Authors: Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, Elisabeth Andr'e

---

Abstract

With the ongoing rise of machine learning, the need for methods for explaining decisions made by artificial intelligence systems is becoming a more and more important topic. Especially for image classification tasks, many state-of-the-art tools to explain such classifiers rely on visual highlighting of important areas of the input data. Contrary, counterfactual explanation systems try to enable a counterfactual reasoning by modifying the input image in a way such that the classifier would have made a different prediction. By doing so, the users of counterfactual explanation systems are equipped with a completely different kind of explanatory information. However, methods for generating realistic counterfactual explanations for image classifiers are still rare. Especially in medical contexts, where relevant information often consists of textural and structural information, high-quality counterfactual images have the potential to give meaningful insights into decision processes. In this work, we present GANterfactual, an approach to generate such counterfactual image explanations based on adversarial image-to-image translation techniques. Additionally, we conduct a user study to evaluate our approach in an exemplary medical use case. Our results show that, in the chosen medical use-case, counterfactual explanations lead to significantly better results regarding mental models, explanation satisfaction, trust, emotions, and self-efficacy than two state-of-the art systems that work with saliency maps, namely LIME and LRP.